Published in final edited form as:
J Biom Biostat. ; 4(1): .

Diagnostic Utility of Gene Expression Profiles

Chengjie Xiong[1,*], **Yan Yan**[1,2], and **Feng Gao**[1]

[1]Division of Biostatistics, Washington University, St. Louis, USA

[2]Department of Surgery, Washington University, St. Louis, USA

Abstract

Two crucial problems arise from a microarray experiment in which the primary objective is to locate differentially expressed genes for the diagnosis of diseases such as cancer and Alzheimer's. The first problem is the detection of a subset of genes which provides an optimum discriminatory power between diseased and normal subjects, and the second problem is the statistical estimation of discriminatory power from the optimum subset of genes between two groups of subjects. We develop a new method to select an optimum subset of discriminatory genes by searching over possible linear combinations of gene expression profiles and locating the one which provides the maximum discriminatory power between two sources of RNA as measured by the area under the receiver operating characteristic (ROC) curve. We further provide an estimate to the optimum discriminatory power between the diseased and the healthy subjects over the selected subsets of genes. The proposed stepwise approach takes in account of the gene-to-gene correlations in the estimation of discriminating power as well as the associated variability and allows the number of genes to be selected based on the increment of the discriminating power. Finally, the proposed methodology is applied to a benchmark microarray experiment and compared to the results obtained through existing approaches in the literature.

Keywords

Confidence interval estimate; Eigenvalue; Eigenvector; Maximum likelihood estimate; Area under curve; Receiver Operating Characteristic (ROC) curve; Fisher's -transformation

Introduction

Transcriptional profiling using microarrays can provide critical information about cellular and tissue expression phenotypes and biological processes. The key statistical quantity in microarray studies is the differential expression of a gene in given experimental conditions. There are multiple sources of variations associated with the observed gene expression level in microarray experiments. The challenge is how to detect the genuinely differential expressions from noisy gene expression data. Both parametric and nonparametric measures

[*]**Corresponding author:** Division of Biostatistics, Campus Box 8067, Washington University, St Louis, MO 63110, USA, Tel: (314)3623635; Fax: (314)362-2693; chengjie@wubios.wustl.edu.

have been proposed to identify discriminatory genes from two experimental conditions. Parametric tests such as *t* test [1] are based on differences of group means, while nonparametric tests such as Wilcoxon's rank sum test are based on differences of rank sums between groups. Parametric tests might not perform well when the underlying distributional assumptions on gene expressions are violated. Nonparametric tests do not assume specific distributional families on gene expressions, but may lack statistical power when certain parametric form of the distribution on gene expressions does exist. Although different microarray data might require different analytic methods based on the specific distributional property of the data, most of these methods in the literature approach the problem by applying them to one gene at a time, and then adjust the p-values from multiple tests by different methods such as the Bonferonni's method. While these approaches are intuitive and relatively easy to implement, they effectively ignore the correlation structure among different genes on the expression data. The correlations on the gene expression data among different genes have been addressed by several authors [2-4]. Techniques based on the analysis of variance (ANOVA) models have been well studied on microarray data [5,6]. Wolfinger et al. [7] presented a statistical approach that allows direct control over the percentage of false positives. Their approach accommodates a wide variety of experimental designs and can simultaneously assess significant differences between multiple types of biological samples. Because their approach is based on a set of general linear mixed models, it provides the possibility of taking into account of the gene to gene correlation in the analysis through various random components in the model. Efron et al. [8] developed a simple empirical Bayes approach to the simultaneous inference problem in gene expression analysis. Their approach produced believable posteriori probabilities of activity differences for each gene, starting with a minimum of a priori assumptions. One of the downside of the empirical Bayes approach is its ad hoc appearance compared to the mathematical certitudes of standard estimation and hypothesis testing theory. Efron and Tibshirani [9] further compared the empirical Bayes approach with the frequentist method of false discovery rates proposed by Benjamini and Hochberg [10]. They pointed out that these two methods were closely related and could be used together to produce sensible simultaneous inferences over a set of possibly correlated genes.

An important emerging medical application domain for microarray gene expression profiling technology is the clinical decision support in the form of diagnosis of a disease as well as the prediction of clinical outcomes in response to a treatment. A number of projects have applied microarray technology to study differences between diseased and healthy tissues [11,12]. These applications have been especially attractive in the management of cancer and infectious disease [12]. Although previous research in this area has established the feasibility of creating accurate models for cancer diagnosis [13], these studies conducted limited experiments in terms of classifiers, gene selection procedures and algorithms, sample sizes, and types of cancers involved. In addition, current approaches in the area provide little information on which classifiers (if any) perform best among the many possible alternatives. There is also another major methodological concern about the problem of overfitting [14,15]. More specifically, the application of gene expression experiments to disease diagnosis involves two fundamental questions. Apart from the question of how to select discriminatory genes from thousands of possible candidates for the use of diagnosis, there is

a crucial question of how to assess the diagnostic accuracy over the many possible classifiers, even from the same set of selected genes. If only two groups of subjects will be compared, the ‘diseased’ and the ‘healthy’, one criterion of discrimination associated with the diagnostic accuracy of the ‘disease’ is assessed through the use of Receiver Operating Characteristic (ROC) curve of gene expressions [16-18]. The ROC curve for the expression of a gene is the plot of sensitivity against (1-specificity). If a gene could perfectly discriminate, the curve would then pass through point (0,1) on the grid [0,1]×[0,1]. The closer an ROC curve comes to this ideal point, the better its discriminating ability. A gene with no discriminating ability will produce a curve that follows the diagonal of the grid. The area under the ROC curve (AUC) [19] has been a particularly popular summary index for the diagnostic accuracy. This area represents the probability that, when the gene expression is observed for a randomly selected individual from the diseased population and a randomly selected individual from the healthy population, the resulting levels will be in the correct order. Both parametric and nonparametric methods have been proposed to estimate the area under a ROC curve [18-22].

When multiple genes are expressed to both healthy and diseased subjects, the resulting ROC curve estimates are correlated due to the correlations on expression levels among different genes. The statistical approach for dealing with multiple correlated diagnostic tests has been largely focused on the comparison of two correlated ROC curves in terms of the discriminating power between the diseased and the healthy populations [23-28]. When searching for differential gene expressions between experimental classes, however, it may not be sufficient to look at each gene in a separate universe. Evaluating combinations of genes might reveal interesting information that will not be discovered otherwise [3]. In another word, we may be interested in not only knowing which gene expression provides the better discriminating power between the diseased subjects and the healthy subjects, but also whether a criterion using a group of genes would give in some sense the best possible discriminating power between these two groups of subjects. The problem of finding the best subset of genes is commonly referred to as the feature subset selection (FSS) problem. Most of the FSS methods begin with evaluating each gene with respect to how well it distinguishes between diseased and the healthy groups, and then rank all genes according to the result and select the top genes as the feature subset to be used. Some also employ a method to remove redundancy in the selected gene set [29]. Linear discriminant analysis (LDA), principal component analysis (PCA), partial least squares regression (PLS) strive to explain most of the variance/covariance structure of the data using linear combinations of the original genes. LDA has often been shown to produce the best classification results. However, it has numerical limitations. In particular, for large data sets with too many correlated predictors, LDA uses too many parameters that are estimated with a high variance. There is therefore a need to either regularize LDA or introduce sparsity in LDA to obtain a parsimonious model. Another limitation of the approaches cited above is the lack of interpretability when dealing with a large number of genes. Ensemble classification methods work on the principle that although a classification algorithm may only be able to produce a model with slightly better accuracy than random guessing, if several such models are produced and combined into an ensemble, their combined accuracy can be greater than any single classifier. Random forest is an algorithm that uses an ensemble of classification trees,

each of which is built using a bootstrap sample of the data, and at each split the candidate set of genes is a random subset of the genes. Thus, random forest uses both bagging (bootstrap aggregation) and random variable selection for tree building. Each tree is unpruned to obtain low-bias trees; at the same time, bagging and random variable selection result in low correlation of the individual trees. The algorithm can be used when the number of genes is much larger than the number of samples. Boosting is also one type of ensemble learning method. Boosting algorithms work by iteratively employing another algorithm known as the base learner to generate a series of models. Initially all samples have equal weights. After the first iteration the accuracy of the model produced is measured and the samples' weights are adjusted so that the weights of misclassified samples are increased while those of correctly classified samples are reduced. At the next iteration the base learner will concentrate on the misclassified samples. A series of models is then produced with the sample weights being adjusted each time. These models are then combined into an ensemble voting. Other authors have reduced the dimensionality by singular value decomposition, and used, for example, the first ten principal components as the feature subset [30,31]. In addition to a large number of supervised and unsupervised methods from the pattern recognition literature, techniques based on linear support vector machines (SVM) are also proposed [32,33]. The methods considering each gene separately potentially miss sets of genes that together allow good discrimination between the experimental classes while each of the genes individually does not. The results by Xiong et al. [31] indicated this may indeed be the case. Bø and Jonassen [3] discussed the benefit when a pair of genes is used to distinguish the healthy versus the diseased tissues and compared their approach to other methods. They showed that gene sets selected based on their method outperform the standard methods, in terms of cross-validation prediction accuracy of the learned classifier. Bayesian approaches were also proposed to identify the optimal nonlinear classifier for diagnosis and the optimal set of genes on which to base that diagnosis [4].

Existing methods in identifying differentially expressed genes and classifying subjects into appropriate classes by gene profiles have used the leave-one-out cross-validation method to appropriately assess the classification accuracy of the proposed classifiers. An appropriate variability measure to the estimated accuracy measure remains a challenge, especially among machine learning techniques that are not based on stochastic models. On the other hand, statistical inferences on diagnostic accuracy based on various statistical models have been well established in the literature. Xiong et al. [34] studied the problem of combining multiple correlated diagnostic tests to provide an optimum test which has the best discriminating power between diseased population and the healthy population. They considered all possible linear combinations of multiple diagnostic tests and searched for the combination that achieved the largest area associated with the ROC curve. Their methodology provided not only an estimate to the optimum diagnostic accuracy but also an appropriate standard error for the estimate.

In this paper we provide a unified approach to select discriminatory genes and to simultaneously estimate their diagnostic accuracy using the expression profiles from these genes. Our approach effectively takes into account the potential correlation on microarray expression data among different genes. More specifically, we provide a stepwise procedure not only to locate an optimum subset of genes which provides the optimum diagnostic

accuracy between the diseased group and the healthy group but also to simultaneously estimate the optimum diagnostic accuracy as measured by the area under ROC curve. Our selection of an optimum subset of genes and the estimation of maximum area under ROC curve are based on the work of Xiong et al. [34]. We will present an algorithm to select an optimum subset of genes to discriminate the diseased and the healthy groups. We will also apply the proposed techniques to a benchmark microarray data set and compare our results with some of those in the literature.

Combining correlated gene expression profiles

We consider that a total of r genes are expressed both in the diseased population and the healthy population. We will follow the convention that higher expressions of each gene are associated with the disease. Let D^+ and D^- denote the diseased group and the healthy group, respectively. Let $(X^1, X^2,....., X^r)^t$ (t stands for the transpose) be the expressions of the r genes for a subject in group D^+, and $(Y^1, Y^2,\ldots, Y^r)^t$ be the expressions of the same r genes for a subject in group D^-. Let l be a r by 1 column vector. Let $U=l^tX$ and $V=l^tY$ be the linear combination of gene expressions in the diseased group and the healthy group, respectively. Notice that the k^{th} gene expression can be obtained by letting $l=(0,0,\ldots,0,1,0,\ldots,0)^t$, where the only 1 occurs at the kth component. Under the assumptions that $(X^1, X^2,....., X^r)^t$ follows a multivariate normal distribution $MVN_r(\mu^+,\sum^+)$ with mean $\mu^+=\left(\mu_1^+,\mu_2^+,\ldots,\mu_r^+\right)^t$ and a positive definite covariance matrix $\sum^+ = (\sigma_{ij}{}^+)_{1\le i,j\le r}$, and that $(Y^1, Y^2,\ldots,$ Yr)t follows another multivariate normal distribution $MVN_r\,(\mu^-,\sum^-)$ with mean $\mu^-=\left(\mu_1^-,\mu_2^-,\ldots,\mu_r^-\right)^t$ and a positive definite covariance matrix $\Sigma^-=\left(\sigma_{ij}^-\right)_{1\le i,j\le r}$, Xiong et al. [34] showed that the maximum area under all possible ROC curves from all possible linear combination expressions is $A=\Phi\left(\sqrt{\lambda_1}\right)$, where λ_1 is the largest eigenvalue of $(\sum^+ + \sum^-)^{-1}(\mu^+ - \mu^-)(\mu^+ - \mu^-)^t$, and Φ is the distribution function of the standard normal distribution. The corresponding eigenvector l_1 provides the optimum linear combination of gene expressions over r different genes. As a special case, when all gene expressions are considered statistically independent, i.e., $\sigma_{ij}^+=\sigma_{ij}^-=0$ for $i \ne j$, the maximum area under the ROC curve over all possible choices of vector l is $\mathrm{A}=\Phi\left(\sqrt{\lambda_1}\right)$, where

$$\lambda_1=\sum_{i=1}^{r}\frac{\left(\mu_i^+ - \mu_i^-\right)^2}{\sigma_{ii}^+ + \sigma_{ii}^-}.$$

The best combination gene expression is when $U=\sum_{i=1}^{r}\left(\frac{\mu_i^+ - \mu_i^-}{\sigma_{ii}^+ + \sigma_{ii}^-}\right)X^i$ for group D^+ and $V=\sum_{i=1}^{r}\left(\frac{\mu_i^+ - \mu_i^-}{\sigma_{ii}^+ + \sigma_{ii}^-}\right)Y^i$ for group D^-.

Suppose that a sample of N subjects undergo tests for determining the presence or absence of the disease. Suppose that it can be determined by means independent of gene expressions

that m of N these individuals truly have the disease, and therefore, $N - m$ subjects are without the disease. Let $X_i=\left(X_i^1, X_i^2, \ldots, X_i^r\right)^t$, $i = 1,2,\ldots,m$, be the values of the r gene expressions for the i^{th} subject in group D^+, and $Y_i=\left(Y_i^1, Y_i^2, \ldots, Y_i^r\right)^t i=1, 2, \ldots, N-m$, be the values of the r gene expressions for the i^{th} subject in group D^-. The maximum likelihood estimators for μ^+ and μ^- are

$$\begin{aligned} \hat{\mu}^+ &= \tfrac{1}{m}\sum_{i=1}^{m} X_i \\ \hat{\mu}^- &= \tfrac{1}{N-m}\sum_{i=1}^{N-m} Y_i, \end{aligned}$$

respectively [35]. The maximum likelihood estimators for $\sum^+$ and are $\sum^-$

$$\begin{aligned} \hat{\Sigma}^+ = & \ \tfrac{1}{m}\left(X_i-\hat{\mu}^+\right)\left(X_i-\hat{\mu}^+\right)^t \\ \hat{\Sigma}^- = & \ \tfrac{1}{N-m}\sum_{i=1}^{N-m}\left(Y_i-\hat{\mu}^-\right)\left(Y_i-\hat{\mu}^-\right)^t, \end{aligned}$$

respectively [35]. Therefore, the estimated best linear combination of the r gene expressions which maximizes the area under ROC curve is achieved when l is the eigenvector $\hat{l}_1$ corresponding to the largest eigenvalue of $\left(\hat{\Sigma}^+ + \hat{\Sigma}^-\right)^{-1}\left(\hat{\mu}^+ - \hat{\mu}^-\right)\left(\hat{\mu}^+ - \hat{\mu}^-\right)^t$, and hence an eigengene. This largest eigenvalue $\hat{\lambda}_1$ is then used to estimate the maximum area under the ROC curve over all possible linear combinations of these r genes as $\hat{A}=\Phi\left(\sqrt{\hat{\lambda}_1}\right)$.

$$\text{Let } \theta=ln\left(\frac{1+A}{1-A}\right) \quad (1)$$

be the Fisher's z-transformation of A. Let $\hat{\theta}=ln\left(1+\hat{A}\right) - ln\left(1-\hat{A}\right)$ be the maximum likelihood estimate of θ. The estimated asymptotic variance of is

$$var\left(\hat{\theta}\right)=\frac{4}{\left(1-\hat{A}^2\right)^2} var\left(\hat{A}\right) \quad (2)$$

The explicit form of $var\left(\hat{A}\right)$ will be mathematically intractable. Xiong et al. [34], however, derived the asymptotic conditional variance of $\hat{A}$, given the estimated eigenvector $\hat{l}_1$. The conditional $var\left(\hat{A}\right)$ is estimated by

$$var\left(\hat{A}\right)=\frac{f^2}{m}\left(1+\frac{m\hat{b}^2}{N-m}+\frac{\hat{a}^2}{2}\right)+\frac{\text{g}\ {}^2\hat{b}^2}{m}\left(\frac{1}{2}+\frac{m}{2\left(N-m\right)}\right)+\frac{fg\hat{a}\hat{b}}{m}, \quad (3)$$

where

$$f = \frac{exp\left[-\hat{a}^2/\left(2+2\hat{b}^2\right)\right]}{\sqrt{2\pi\left(1+\hat{b}^2\right)}},$$
$$g = -\frac{\hat{a}\hat{b}exp\left[-\hat{a}^2/\left(2+2\hat{b}^2\right)\right]}{\sqrt{2\pi\left(1+\hat{b}^2\right)^3}},$$
$$\hat{a} = \frac{l_1^t\hat{\mu}^+ - l_1^t\hat{\mu}^-}{\sqrt{l_1^t\hat{\Sigma}^+ l_1}},$$
$$\hat{b} = \frac{\sqrt{l_1^t\hat{\Sigma}^- l_1}}{\sqrt{l_1^t\hat{\Sigma}^+ l_1}}.$$

This leads to an asymptotic 100(1–α)% (0 <α< 1) conditional confidence interval to the largest area $A=\Phi\left(\sqrt{\lambda_1}\right)$ over all possible linear combination tests as $\left[\hat{A}_1, \hat{A}_2\right]$, where

$$\hat{A}_1 = z^{-1}\left[\hat{\theta} - z_{\alpha/2}\sqrt{var\left(\hat{\theta}\right)}\right],$$
$$\hat{A}_2 = z^{-1}\left[\hat{\theta} + z_{\alpha/2}\sqrt{var\left(\hat{\theta}\right)}\right], \quad (4)$$

and z^{-1} is the inverse of z :

$$z^{-1}(\theta) = \frac{exp(\theta) - 1}{exp(\theta) + 1}.$$

A simulation study by Xiong et al. [34] showed that under the assumption of multivariate normal distribution on the gene expressions, the proposed confidence interval estimate performed well in achieving the nominal confidence level, even when the sample sizes are relatively small.

Selecting an optimum subset of differentially expressed genes

Microarray experiments typically involve thousands of genes. Most times only a small proportion of these genes are genuinely differentially expressed between the diseased and the healthy subjects. We now propose a stepwise procedure to select a set of optimum s genes out of a total of r genes which in certain sense optimally discriminate the diseased from the healthy groups.

Step 1

select the single g_1 gene that maximizes the area under the ROC curve out of a total of r genes. For the k^{th} gene, $1 \le k \le r$, the corresponding ROC curve is defined by the function

$$f_k(x) = \Phi\left[a_k + b_k\Phi^{-1}(x)\right],$$

where

$$a_k = \frac{\mu_k^+ - \mu_k^-}{\sqrt{\sigma_{kk}^+}},$$
$$b_k = \sqrt{\frac{\sigma_{kk}^-}{\sigma_{kk}^+}},$$

and Φ is the distribution function of the standard normal distribution, x is one minus the specificity. The area under the ROC curve (England, 1988) is given by

$$A_k = \int_0^1 f_k(x)\,dx = \Phi\left(\frac{a_k}{\sqrt{1+b_k^2}}\right).$$

The gene g_1 chosen at this step is the one such that A_k is maximized over the choice of k. We denote the maximum area by $A(g_1)$.

Step 2

Now that gene $g_1, g_2, \ldots, g_{s-1}$ are chosen, for each gene g other than $g_1, g_2, \ldots, g_{s-1}$, we first consider all possible linear combinations among genes $g_1, g_2, \ldots, g_{s-1}$ and gene g, and locate the maximum area under the ROC curve over all possible s-gene linear combinations of gene expressions. This process can be done by using the results described above when $r=s$. Denote the maximum area under the ROC curve over all possible s-gene linear combinations between gene $g_1, g_2, \ldots, g_{s-1}$ and gene g by $A(g_1, g_2, \ldots, g^{s-1}, g)$. We then choose gene such that $A(g_1, g_2, \ldots, g_{s-1}, g)$ is maximized over the choice of g, and denote the maximum area by $A(g_1, g_2, \ldots, g_{s-1}, g_s)$.

The above stepwise process defines an optimum set of s genes which provides in some sense the optimum discrimination power between the diseased and the healthy groups. Notice that the genes chosen by the previous steps are always kept when an additional gene is added at the next step. Therefore, it is always true that, when an additional gene is added to the list in the process, the maximum area under the ROC curves is increasing, i.e., $A(g_1) \le A(g_1, g_2) \le \ldots \le A(g_1, g_2, \ldots, g_{s-1}, g_s)$. Notice also that the entire process does not have to begin with the optimum single gene that maximizes the area under the ROC curve out of a total of r genes. Sometimes certain gene expressions have been proven to discriminate well the diseased group from the healthy group. If the objective of a microarray experiment is to identify other novel discriminatory genes, then there is no reason to include the well known discriminatory genes in the search process, in which case, we would be selecting a set of optimum genes from a subset of unknown genes.

Step 3 (Stopping Rule for the Selection of an Optimum Subset of Genes)

Our proposed stepwise procedure for selecting an optimum subset of genes could work in two ways: when the number of genes in the optimum subset is or is not prespecified. In the latter case, a statistical stopping rule is needed to terminate the stepwise procedure. We propose here a statistical stopping rule based on the estimate to the diagnostic accuracy with

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the selected genes. At the s^{th} step, a total of s genes are selected. Suppose that a sample of total N subjects is used m for the microarray experiment for which of these N individuals truly have the disease, and $N - m$ subjects are without the disease. The maximum likelihood estimates to mean vectors and covariance matrices of s gene expressions can be obtained as described in the Section above (Combining Correlated Gene Expression Profiles) for both subject groups. The maximum area under the ROC curve can be readily estimated by the confidence interval (4). We point out, however, when the number of genes is at least as large as the sample size, the maximum likelihood estimates to the covariance matrix will only be semi-positive definite but not positive definite, and the process of gene selection will not be possible when the number of genes to be selected is large. In fact, with a sample size m in the diseased group and a sample size $N - m$ in the healthy group, the maximum number of optimum genes that can be detected by the above process is the maximum of m^{-1} and N-m^{-1} [35]. This is due to the fact that the computation of the optimum linear combination of gene expressions from a set of multiple genes requires the existence of with $\left(\hat{\Sigma}^{+}+\hat{\Sigma}^{-}\right)^{-1}$ probability 1, i.e., at least one of the two estimated covariance matrices is positive definite.

In order to provide a statistical stopping rule for the stepwise procedure, we compare the optimum diagnostic accuracy obtained from the s^{th} step and that from the following $(s+1)^{\text{th}}$ step. Let

$$\begin{aligned}\theta_s= & \quad ln\left(\frac{1+A(g_1,g_2,\ldots,g_{s-1},g_s)}{1-A(g_1,g_2,\ldots,g_{s-1},g_s)}\right),\\ \theta_{s+1} & =ln\left(\frac{1+A(g_1,g_2,\ldots,g_s,g_{s+1})}{1-A(g_1,g_2,\ldots,g_s,g_{s+1})}\right).\end{aligned}$$

Let $\hat{\theta}_s$ and $\hat{\theta}_{s+1}$ be the maximum likelihood estimate of θ_s and θ_{s+1}, respectively. We test the null hypothesis $H_0 : \theta_s = \theta_{s+1}$ against the alternative hypothesis $H_a : \theta_s < \theta_{s+1}$. An asymptotic size $\alpha(0 < \alpha < 1)$ test rejects the null hypothesis when $z > z_\alpha$, where z_α is the upper 100α% percentile of the standard normal distribution, and

$$\begin{aligned}z & =\frac{\hat{\theta}_{s+1}-\hat{\theta}_s}{\sqrt{var\left(\hat{\theta}_{s+1}-\hat{\theta}_s\right)}},\\ & var\left(\hat{\theta}_{s+1}-\hat{\theta}_s\right)=var\left(\hat{\theta}_{s+1}\right)+var\left(\hat{\theta}_s\right)-2cov\left(\hat{\theta}_{s+1},\hat{\theta}_s\right).\end{aligned}$$

The variance of $\hat{\theta}_i, i=s, s+1$, is

$$var\left(\hat{\theta}_i\right)=\frac{4}{\left[1-\hat{A}^2\left(g_1,g_2,\ldots,g_i\right)\right]^2}var\left(\hat{A}\left(g_1,g_2,\ldots,g_i\right)\right),$$

where $var\left(\hat{A}\left(g_1,g_2,\ldots,g_i\right)\right)(i=s, s+1)$ is given by Equation (3) when applied to the optimum linear combination of genes $g_1, g_2,\ldots, g_i$. Further,

$$cov\left(\hat{\theta}_{s+1},\hat{\theta}_{s}\right)=\frac{4cov\left[\hat{A}\left(g_{1},g_{2},\ldots,g_{s}\right),\hat{A}\left(g_{1},g_{2},\ldots,g_{s},g_{s+1}\right)\right]}{\left[1-\hat{A}^{2}\left(g_{1},g_{2},\ldots,g_{s}\right)\right]\left[1-\hat{A}^{2}\left(g_{1},g_{2},\ldots,g_{s},g_{s+1}\right)\right]},$$

where $cov\left[\hat{A}\left(g_{1},g_{2},\ldots,g_{s}\right),\hat{A}\left(g_{1},g_{2},\ldots,g_{s},g_{s+1}\right)\right]$ can be computed by Formula (10) through Formula (14) of Obuchowski and McClish [36] when applied to the correlated optimum linear combination of genes $g_1,g_2,\ldots,g_s$ and the optimum linear combination of genes $g_1,g_2,\ldots,g_s,g_{s+1}$. Notice that all the variances and covariance used above are conditioning on the optimum linear combination of genes $g_1,g_2,\ldots,g_i$ for $i=s$, $s+1$. We propose to perform the above test at each step and to stop the stepwise procedure for selecting an optimum subset of genes at step s when the null hypothesis $H_0 : \theta_s = \theta_{s+1}$ is not rejected against the alternative hypothesis $H_a{:}\,\theta_s<\theta_{s+1}$. Intuitively, the proposed stopping rule stops the stepwise procedure for selecting an optimum subset of genes when adding another extra gene does not appreciably increase the optimum diagnostic accuracy measure. We point out that, to avoid model over fitting that could lead to too many genes to be chosen, the control of Type I error rate is very important for the repeated tests on the sequential hypotheses in the comparisons between the current AUC and the next AUC. One way to implement this is to assess the joint distribution of all sequentially estimated θ *'s* .

These sequential estimates are correlated and follow an asymptotically multivariate normal distribution (conditional on the estimated eigenvectors). Hence, some types of alpha-spending procedures similar to those proposed in group sequential clinical trials [37] can be developed to control the overall Type I error rate in the repeated tests. The details of these procedures are out of the scope of the current manuscript, and will be developed in our subsequent work. We also point out that the proposed stopping rule should only be considered as a method to terminate the search process of genes when the optimum diagnostic accuracy does not increase measurably. This stopping rule does not imply, however, that genes out of the chosen optimum subset lack the discriminatory ability to differentiate the diseased subjects from the healthy subjects. In fact, after an optimum subset of genes is located first, the stepwise procedure might be employed again to search for another optimum subset of genes from the pool of genes left after the first search process. This latter process might very well reveal more discriminatory genes, and result in multiple eigengenes from different optimum subsets of genes which can be assessed based on their varying degree of discriminating power between the diseased group and the healthy group. In fact, the proposed methodology can also be applied to combine the multiple eigengenes to further improve the diagnostic accuracy between the diseased and healthy groups.

There are several important features of our proposed approach for detecting differentially expressed genes. First, unlike many reports in the literature which are based on various methods of statistical hypothesis tests, we address the question by estimating a discrimination index (e.g., the maximum area under ROC curve) which measures the optimum discriminating power from a set of genes. We also provide an estimate to the variance associated with the estimated optimum discrimination index. Because the discrimination index is the maximum area under ROC curves from all possible linear

combinations of gene expressions from the set of genes, this index represents the maximum probability that, when multiple gene expressions are observed for a randomly selected individual from the diseased population and a randomly selected individual from the healthy population, the best linearly combined gene expressions over multiple genes will be in the correct order. Notice that this discrimination index is a well defined distributional parameter which is uniquely determined by the joint distributions of gene expressions over multiple genes between the diseased and healthy subjects. Another advantage of our proposed approach is that the proposed discrimination index can be readily and consistently estimated with a large sample size through the method of maximum likelihood using the entire sample available, which does not require the validation and bias adjustment based on methods such as the leave-one-out cross validation. Whereas our approach does estimate the diagnostic accuracy from the same sample that is used to select the genes, the (conditional) variance to the estimated diagnostic accuracy, conditioning on the estimated optimum linear combination, is valid. We point out that, however, due to the potential selection bias, the assessment of the unconditional variance of the diagnostic accuracy of selected genes must be obtained through a cross validation or bootstrap procedure that is external to the gene selection process [38]. Second, as pointed out by Pan et al. [39], the variances of gene expressions among different subject groups are likely dependent on the mean expression level, and our proposed approach accounts for this dependence by allowing different covariance matrices between the diseased group and the healthy group. Third, although many authors analyzed microarray experiment data by assuming the statistical independence on the expression levels among different genes [1], we would argue that there are many possible ways that correlations exist on the expression levels among different genes. Apart from the possible biological correlation among different genes, the fact that multiple expression data are obtained from the same microarray and the same study subject might introduce possible generic and spatial correlations among different gene expressions. Our proposed approach takes into account of these possible correlations and the possible disparity on these correlations by assuming two different unstructured covariance matrices for the diseased group and the healthy group. On the other hand, we also point out that, when we are willing to assume a structured covariance structure for the gene expression level among the entire set of genes, the proposed stepwise procedure for the selection of optimum genes could still work as long as the maximum likelihood estimates to the structured covariance matrices are appropriately adjusted. One likely structured covariance structure is a form called the compound symmetry [40]. A covariance matrix is of compound symmetry form if the variances of the gene expressions are the same across different genes, and the correlations of the gene expressions are also the same between any two different genes. The maximum likelihood estimate to the covariance matrix of compound symmetry form from an independent sample of multivariate normal distribution is described by Milliken and Johnson [40]. In addition, when such a covariance matrix is assumed, it satisfies the so-called Huynh-Feldt condition [41] on repeated measures from different genes on the same array, and the microarray gene expression data can be readily analyzed by the classical analysis of variance models under the split-plot design as the resulting *F*-test for the interactive effect between subject groups and gene factor are still valid [41].

Application and comparison to existing methods

To evaluate the performance of our proposed methodology, we apply our methods to a well known benchmark public dataset. The Alon et al. [11] data set consists of 62 samples, of which 22 are normal and 40 are colon tumor tissues. Gene-expression levels were measured using Affymetrix oligonucleotide arrays complementary to more than 6,500 genes. Published along with the Alon et al. [11] paper was a dataset containing expression levels for the 2,000 genes with the highest minimal intensity across the samples, and this is the dataset we study in this paper. Before analysis, we carried out the following preprocessing steps on the dataset: first apply the base 10 logarithms; and then for each gene, subtract the mean and divide by the standard deviation.

Using our proposed method, we first searched for an optimum subset of 15 genes to differentially discriminate the two experimental conditions based on the optimum AUC as the measure of diagnostic accuracy. These optimum 15 genes are presented in the first column of table 1 in the order of entering the subset from our stepwise procedure. For the purpose of comparison, the top 15 genes in the order of individual ranking based on evaluating each gene separately with a two sample t-statistic from the entire pool of genes are also presented in the third column of table 1. These top 15 genes based on individual ranking were also reported by [3]. In addition, we report at each row of table 1 the z-transformation of the estimated maximum AUC from all possible linear combination of the genes up to the row along with the corresponding estimated standard error. Notice that many genes selected by our proposed method are not from the top 15 genes based on the individual ranking, indicating that genes which discriminate well as a whole might not necessarily be those that discriminate well individually. It is also interesting to observe that the maximum z-transformation of the AUC over all possible linear combinations from our proposed method are consistently larger than those based on the genes with high individual rankings, indicating that genes with high individual rankings as a whole might not necessarily jointly provide the best discriminating power between two experimental conditions.

The same data set was also studied by Bø and Jonassen [3] in which pairs of genes were used to distinguish two experimental conditions. The approach by Bø and Jonassen [3] was based on a pair score which was essentially the two sample t-statistic of the projected points on the diagonal linear discriminant axis using only two genes [42]. The pairs of genes were then ranked based on the magnitude of the pair score. Although the ranking of pairs of genes on a t statistic in Bø and Jonassen [3] is geometrically appealing, it lacks a direct connection with the measurement of diagnostic accuracy. Notice also that the approach by Bø and Jonassen [3] is very computationally intensive because it was based on the all-pairs procedure, i.e., all possible pairs of genes from the entire pool of genes were considered. More specifically, given pair scores for all pairs, they selected the top-ranked disjoint pairs in a greedy manner. First, the pair with highest pair score was selected, then all pairs containing any of these two genes were removed from the list. Finally the highest-scoring pair from the remaining list was chosen, and so on. When applied to search for the optimum pairs of differentially expressed genes, our approach differs from that of Bø and Jonassen [3] in a couple of fronts. First, our approach not only provides a direct connection with the

measurement of diagnostic accuracy as it is based on the area under ROC curve, but also offers a measure of the optimum ability a pair of genes possesses to differentiate the experimental conditions (e.g., maximum area under ROC curves from all the possible linear combinations of two gene expression profiles in each pair). Second, for the selection of each pair of genes, our stepwise procedure begins with the most highly individually ranked gene from the pool of available genes, and then search for the other gene that provides the maximum area under the ROC curve over all possible linear combinations of two gene expression profiles. Therefore, our approach is not a greedy search procedure, which is much less computationally intensive than the all pairs greedy search procedure of Bø and Jonassen [3]. We first applied our proposed methodology to search for the best pairs of genes differentiating the experimental conditions and then compared our results with the results from Bø and Jonassen [3]. These comparisons were based on the estimated maximum area under the corresponding ROC curve over all possible linear combinations of gene expressions from each pair. More specifically, we first repeatedly applied our stepwise procedure to search for the optimum top 25 pairs of genes and then compared our pairs with the top 25 ranked pairs as reported by Bø and Jonassen [3]. In our repeated search for the top 25 ranked pairs, the next pair was always selected after the previously selected pairs of genes were excluded. By removing the already selected genes from the gene pool, we did not take the risk that one exceptionally differentially expressed gene can drag along with several mediocre companion genes. If such a highly differentially expressed gene was left in the gene set, it would probably be responsible for many of the top-ranked pairs. Therefore, by removing selected genes from the gene pool, a highly differentially expressed gene would only cause its best available companion to join it in the set of selected genes. Table 2 presents both the top-ranked 25 pairs of genes for the class separation using all pairs ranking method from Bø and Jonassen [3] (column 1) and the top-ranked 25 pairs of genes using our proposed methodology (column 3). For each pair of genes in table 2, we also computed the z-transformation of the estimated maximum AUC (zAUC) from all possible linear combinations of the pair of genes. An estimated standard error for each estimated zAUC is also presented for each pair of genes in table 2. The pairs of genes are rearranged based on the z-transformation of the estimated maximum AUC from all possible linear combinations of the pair of genes in table 2.

Table 2 demonstrates some interesting comparisons. First, the 25 pairs of top ranked genes based on our proposed methodology differ widely from those based on the all pairs ranking method from Bø and Jonassen [3]. When these different pairs of genes are compared on the same footing by the z-transformation of the estimated maximum area under ROC curve (AUC) from all possible linear combinations of a pair of genes, our proposed methodology provides a better diagnostic accuracy for the top 7 pairs compared to the top 7 pairs from [3], although the differences are unlikely to be statistically significant due to the lack of statistical power from the limited sample size. The total of these 14 pairs of genes are likely to be the most important pairs in terms of the diagnostic accuracy as they provide a minimum estimated optimum area under the ROC curve of 95.92%, indicating that, when a random pair of patients with one from the normal group and the other from the tumor group, these two patients can be correctly diagnosed using the expression profiles of these 14 pairs of genes with an estimated minimum probability of 95.92%. It is also interesting to observe

that our proposed methodology offers slightly less diagnostic accuracy for some other pairs of genes in table 2 compared to these from [3], although the differences are unlikely to be statistically significant. This could be explained by the fact that all possible pairs of genes from the entire pool of genes were searched and compared in a greedy manner by Bø and Jonassen [3], while our proposed method is much less computationally intensive by only searching for 1 gene from the available pool of genes in an optimum manner. Therefore, our proposed method offers a much less computationally intensive method to select subsets of genes which are comparable in the measure of diagnostic accuracy to these from Bø and Jonassen [3]. Table 3 provides Annotations of genes appeared in table 1 and table 2.

Discussion

A fundamentally important question in microarray experiments associated with a disease is which genes are involved and which genes are possible marker genes for the disease. We proposed a methodology to not only locate an optimum subset of genes which provides the maximum discriminating power between the diseased subjects and the healthy subjects but also simultaneously estimate the optimum discriminating power as measured by the area under ROC curves. Another very important feature of the proposed methodology is the incorporation of gene by gene correlation arising from the expression profiles. In our stepwise algorithm of locating the optimum subset of differentially expressed genes, we not only presented the MLE for the optimum area under the ROC curves at each step, but also provided the corresponding estimated standard error and a confidence interval estimate to this optimum measure of diagnostic accuracy which performed well with the typical sample sizes used in microarray experiments [34]. The latter was especially important in the statistical assessment of diagnostic accuracy because it provides the degree of variation in the point estimate to the optimum diagnostic accuracy. This also contrasted with most of the machine learning based techniques reported in the literature which have largely ignored the variation in the assessment of classification accuracy when the estimates based on leave-one-out cross validation were reported without the associated estimated standard errors. Our proposed methodology may therefore provide further insight into the biological mechanisms behind a disease. In fact, the application of our proposed methodology to an existing benchmark data set revealed some genes that are not obviously good discriminators alone when each gene was evaluated individually, but discriminated well when combined with other genes and when the gene by gene correlations are taken into account. Therefore, our proposed methodology could help discover interesting genes that could be overlooked in the traditional approaches. Notice that our proposed methodology was based on a discrimination index which measures the optimum discriminating power from a set of genes and is uniquely decided by the joint distributions of gene expressions over multiple genes between the diseased and healthy subjects. Because the MLE offers a valid and consistent estimate with large sample size, our proposed methodology does not require the cross-validation based on methods such as the leave-one-out approach. On the other hand, such cross-validation techniques will be very important and necessary in the assessment of diagnostic accuracy when an estimated optimum linear combination associated with our estimated optimum area under ROC curve is to be used with a specified threshold for disease diagnosis.

When applied to a benchmark public dataset of microarray experiments as reported in table 1, our proposed methodology of locating the optimum subset of 15 genes outperforms the 15 most highly individually ranked genes (based on a two-sample *t* test), most times by a large margin, in the measure of *z*-transformation of the maximum area under ROC curves over all possible linear combinations of gene expression profiles. In our comparison in terms of diagnostic accuracy with the top ranked 25 pairs of genes from Bø and Jonassen [3] in table 2, we have shown that our proposed methodology provides a better diagnostic accuracy for the top 7 pairs. Given the fact that our proposed method in the search of top ranked pairs of genes is much less computationally intensive as compared to that by Bø and Jonassen [3] where all possible pairs of genes from the entire pool of genes were searched and compared in a greedy manner, the results of table 2 seems to indicate that our proposed method offers a much less computationally intensive method to select subsets of genes which are comparable in the measure of diagnostic accuracy to these chosen in a greedy search.

We do not try to claim that a single application of our proposed methodology will enable us to find all discriminatory genes associated with a disease, as there may be relevant genes that are biologically significant by themselves but may not appear in the optimum subset of genes chosen. What we have demonstrated is, however, a statistical methodology that combines the correlated gene expression profiles to achieve the optimum diagnostic accuracy as measured by the area under ROC curves. We believe that our approach is a step in the right direction to appropriately combine the expression profiles from multiple genes for the benefit of accurate disease diagnosis. On the other hand, there could well be several different subsets of differentially expressed genes which will all provide excellent discriminating ability between the diseased subjects and normal subjects. The repeated use of our proposed methodology will help to identify these multiple subsets of genes. One limitation of our proposed methodology is that it can only be applied when two experimental conditions are used in microarray experiments. How the proposed approach can be generalized into the situation when multiple tumor types are used in a microarray experiment remains an open question which will be tackled by our subsequent research.

Acknowledgments

The authors thank the editor and reviewers for their constructive comments that have improved the manuscript considerably. This study was supported by National Institute of Health (NIH) grant R01 AG029672 and R01 AG034119 for Chengjie Xiong. This study was also partly supported by the NIH grant P50 AG05681, P01 AG03991, P01AG26276, and U01 AG032438 for Chengjie Xiong.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286:531–537. [PubMed: 10521349]
2. Simon R. Using DNA microarrays for diagnostic and prognostic prediction. Expert Rev Mol Diagn. 2003; 3:587–595. [PubMed: 14510179]
3. Bø T, Jonassen I. New feature subset selection procedures for classification of expression profiles. Genome Biol. 2002; 3:Research0017. [PubMed: 11983058]
4. Krishnapuram B, Carin L, Hartemink AJ. Joint classifier and feature optimization for comprehensive cancer diagnosis using gene expression data. J Comput Biol. 2004; 11:227–242. [PubMed: 15285890]

5. Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarrays. Genet Res. 2001; 77:123–128. [PubMed: 11355567]
6. Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, et al. Statistical analysis of a gene expression microarray experiment with replication. Statistica Sinica. 2002; 12:203–217.
7. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, et al. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol. 2001; 8:625–637. [PubMed: 11747616]
8. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc. 2001; 96:1151–1160.
9. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol. 2002; 23:70–86. [PubMed: 12112249]
10. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995; 57:289–300.
11. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A. 1999; 96:6745–6750. [PubMed: 10359783]
12. Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: and empirical assessment. Lancet. 2003; 362:1439–44. [PubMed: 14602436]
13. Lee Y, Lee CK. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics. 2003; 19:1132–1139. [PubMed: 12801874]
14. Reunanen J. Overfitting in making comparisons between variables selection methods. J Mach Learn Res. 2003; 3:1371–1382.
15. Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. Stat Med. 2000; 19:541–561. [PubMed: 10694735]
16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]
17. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988; 240:1285–1293. [PubMed: 3287615]
18. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8:283–298. [PubMed: 112681]
19. Swets, JA.; Pickett, RM. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York; Academic Press: 1982.
20. DeLong ER, Vernon WB, Bollinger RR. Sensitivity and specificity of a monitoring test. Biometrics. 1985; 41:947–958. [PubMed: 3913467]
21. Ma G, Hall WJ. Confidence bands for receiver operating characteristic curves. Med Decis Making. 1993; 13:191–197. [PubMed: 8412547]
22. Dorfman DD, Alf E Jr. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating-method data. J Math psychol. 1969; 6:487–496.
23. Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC-curve estimates obtained from partially-paired data sets. Med Decis Making. 1998; 18:110–121. [PubMed: 9456215]
24. Metz, CE.; Wang, P-L.; Kronman, HB. Information Processing Medical imaging VIII. Springer; Netherlands: 1984. A new approach for testing the significance of differences between ROC curves measured from correlated data.
25. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44:837–845. [PubMed: 3203132]
26. Venkatraman ES, Begg CB. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika. 1996; 83:835–848.
27. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika. 1989; 76:585–592.
28. Zhou, X-H.; Obuchowski, NA.; McClish, DK. Statistical Methods in Diagnostic Medicine. Wiley-Interscience; 2002.

29. Xing, EP.; Jordan, MI.; Karp, R. Feature selection for high-dimensional genomic microarray data. Proceedings of Eighteenth International Conference on Machine Learning; San Francisco. 2001.
30. Khan J, Wei JS, Rigner M, Saal LH, Ladany IM, Wetsermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine. 2001; 7:673–679.
31. Xiong M, Jin L, Li W, Boerwinkle E. Computational methods for gene expression-based tumor classification. BioTechniques. 2000; 29:1264–1268. [PubMed: 11126130]
32. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000; 16:906–914. [PubMed: 11120680]
33. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002; 46:389–422.
34. Xiong C, McKeel DW Jr, Miller JP, Morris JC. Combining correlated diagnostic tests---application to neuropathologic diagnosis of Alzheimer's disease. Med Decis Making. 2004; 24:659–669. [PubMed: 15534346]
35. Graybill, FA. Theory and Application of the Linear Model. Duxbury; North Scituate: 1976.
36. Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Stat Med. 1997; 16:1529–1542. [PubMed: 9249923]
37. Jennison, C.; Turnbull, BW. Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC; Boca Raton: 2000.
38. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA. 2002; 99:6562–6566. [PubMed: 11983868]
39. Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biology 3: research0022. 2002
40. Milliken, GA.; Johnson, DE. Analysis of Messy Data, Volume 1: Designed Experiments. New York; Chapman & Hall/CRC: 1992.
41. Huynh H, Feldt LS. Conditions under which mean square ratios in repeated measures designs have exact F-distributions. J Am Stat Assoc. 1970; 65:1582.
42. Mardia, KV.; Kent, JT.; Bibby, JM. Multivariate Analysis. Academic Press; London: 1979.

Table 1

Optimum 15 genes for colon tumor/normal class discrimination

Gene ID (in the order of entering the subset from the proposed method)	zAUC to column 1 (Standard error)	Gene ID (in the order of individual ranking)	zAUC to column 3 (Standard error)
R87126	2.76 (0.40)	R87126	2.76 (0.40)
X55715	4.09 (0.83)	M63391	2.82 (0.56)
T86444	4.61 (0.94)	M26383	3.12 (0.64)
T87527	5.24 (1.07)	H08393	4.41 (0.97)
R62549	5.85 (1.21)	X12671	4.75 (1.03)
L37792	6.46 (1.36)	R36977	4.77 (1.03)
H08393	7.90 (1.80)	J02854	4.78 (1.04)
H16096	9.63 (2.21)	J05032	4.79 (1.04)
X52151	10.85 (2.59)	Z50753	5.06 (1.14)
R88740	12.19 (3.04)	M76378	5.08 (1.11)
D14812	15.07 (3.83)	M22382	5.23 (1.16)
X93510	17.48 (4.52)	X63629	5.42 (1.24)
J04794	21.22 (5.41)	M76378	5.57 (1.29)
U09564	25.62 (6.04)	H43887	5.60 (1.30)
R80966	30.04 (7.30)	M36634	5.61 (1.29)

Table 2

Comparison of top-ranked 25 pairs of genes for colon tumor/normal class discrimination (zAUC= *z*-transformation of the maximum area under ROC curve)

Pair Rank	Gene ID from [3]	zAUC (standard error)	Gene ID (the proposed method)	zAUC (standard error)
1	J05032		Z50753	
1	U19969	4.27 (0.79)	H09719	4.43 (0.97)
2	X86693		X12671	
2	D14812	4.20 (0.78)	X12369	4.33 (0.89)
3	M63391		J05032	
3	H08393	4.19 (0.90)	U19969	4.27 (0.79)
4	R87126		H08393	
4	X63629	4.05 (0.84)	M63391	4.19 (0.90)
5	M36634		J02854	
5	H11084	4.05 (0.81)	T57882	4.16 (0.91)
6	X12671		R87126	
6	Z50753	4.04 (0.90)	X55715	4.09 (0.83)
7	H06524		R36977	
7	U22055	3.87 (0.73)	H20709	3.90 (0.70)
8	M76378		M22382	
8	T62947	3.86 (0.70)	R55310	3.83 (0.80)
9	J02854		M76378	
9	R54097	3.85 (0.77)	T62947	3.77 (0.72)
10	Z48541		X14958	
10	D25217	3.67 (0.62)	L06895	3.61 (0.65)
11	D21261		U30825	
11	H20709	3.57 (0.82)	T62878	3.52 (0.61)
12	T90280		X63629	
12	T51534	3.48 (0.62)	M36634	3.50 (0.64)
13	T92451		T71025	
13	U09587	3.46 (0.66)	L11706	3.45 (0.68)
14	H09719		U09564	
14	L07648	3.45 (0.73)	T64467	3.43 (0.66)
15	T51023		R84411	
15	D31716	3.45 (0.61)	M92287	3.37 (0.63)
16	T71025		M76378	
16	L11706	3.45 (0.67)	D00860	3.36 (0.63)
17	X12369		M26697	
17	R98842	3.44 (0.70)	T47424	3.34 (0.60)
18	X14958		T86749	

Pair Rank	Gene ID from [3]	zAUC (standard error)	Gene ID (the proposed method)	zAUC (standard error)
18	X87159	3.43 (0.66)	M74491	3.33 (0.71)
19	J04102		M26383	
19	U14631	3.34 (0.70)	T47377	3.30 (0.59)
20	M76378		X54942	
20	D00860	3.30 (0.65)	R44301	3.28 (0.70)
21	M26383		H43887	
21	T47377	3.29 (0.59)	U26312	3.28 (0.55)
22	X54942		M76378	
22	R44301	3.28 (0.69)	T56604	3.27 (0.58)
23	M76378		T86473	
23	T56604	3.26 (0.58)	D25217	3.20 (0.53)
24	R96357		H40095	
24	R46753	3.26 (0.66)	H06524	3.03 (0.52)
25	T63133		T95018	
25	T61661	2.66 (0.45)	Z49269	2.85 (0.53)

NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 3

Annotations of genes appeared in Table 1 and Table 2

Gene ID	Annotations
H20709	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (HU-MAN).
T95018	40S RIBOSOMAL PROTEIN S18 (Homo sapiens).
T61661	PROFILIN I (HUMAN).
T71025	Human (HUMAN).
T51534	CYSTATIN C PRECURSOR (HUMAN).
X55715	Human Hums3 mRNA for 40S ribosomal protein s3.
T62878	CYTOCHROME C OXIDASE POLYPEPTIDE IV PRECURSOR (HUMAN);.
D25217	Human mRNA (KIAA0027) for ORF, partial cds.
M26697	Human nucleolar protein (B23) mRNA, complete cds.
D21261	SM22-ALPHA HOMOLOG (HUMAN);.
T51023	HEAT SHOCK PROTEIN HSP 90-BETA (HUMAN).
T63133	THYMOSIN BETA-10 (HUMAN);.
T64467	P33477 ANNEXIN XI ;.
T47424	INSULIN RECEPTOR SUBSTRATE-1 (Homo sapiens)
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6.
M63391	Human desmin gene, complete cds.
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6.
T87527	HEAT SHOCK PROTEIN HSP 84 (Mus musculus)
T57882	MYOSIN HEAVY CHAIN, NONMUSCLE TYPE A (Homo sapiens)
X14958	Human hmgI mRNA for high mobility group protein Y.
Z50753	H.sapiens mRNA for GCAP-II/uroguanylin precursor.
R80966	CLATHRIN LIGHT CHAIN B (HUMAN).
U30825	Human splicing factor SRp30c mRNA, complete cds.
X87159	H.sapiens mRNA for beta subunit of epithelial amiloride-sensitive sodium channel.
M74491	Human ADP-ribosylation factor 3 mRNA, complete cds.
R87126	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
M22382	MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN).
T56604	TUBULIN BETA CHAIN (Haliotis discus)
R46753	CYCLIN-DEPENDENT KINASE INHIBITOR 1 (Homo sapiens)
D14812	Human mRNA for ORF, complete cds.
J04794	Human aldehyde reductase mRNA, complete cds.
L06895	Homo sapiens antagonizer of myc transcriptional activity (Mad) mRNA, complete cds.
X12671	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1.
Z48541	H.sapiens mRNA for protein tyrosine phosphatase.

Gene ID	Annotations
X12369	TROPOMYOSIN ALPHA CHAIN, SMOOTH MUSCLE (HUMAN).
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6.
H40095	MACROPHAGE MIGRATION INHIBITORY FACTOR (HUMAN).
R88740	ATP SYNTHASE COUPLING FACTOR 6, MITOCHONDRIAL PRECURSOR (HUMAN).;.
Z49269	H.sapiens gene for chemokine HCC-1.
T92451	TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE (HUMAN).
H43887	COMPLEMENT FACTOR D PRECURSOR (Homo sapiens)
T86473	NUCLEOSIDE DIPHOSPHATE KINASE A (HUMAN)
T90280	RIBOPHORIN II PRECURSOR (HUMAN).
R36977	P03001 TRANSCRIPTION FACTOR IIIA.
U09587	Human glycyl-tRNA synthetase mRNA, complete cds.
U09564	Human serine kinase mRNA, complete cds.
R98842	PROTHYMOSIN ALPHA (Homo sapiens)
D31716	Human mRNA for GC box bindig protein, complete cds.
R84411	SMALL NUCLEAR RIBONUCLEOPROTEIN ASSOCIATED PROTEINS B AND B’ (HUMAN);
R96357	POLYADENYLATE-BINDING PROTEIN (Xenopus laevis)
R55310	S36390 MITOCHONDRIAL PROCESSING PEPTIDASE .
R62549	PUTATIVE SERINE/THREONINE-PROTEIN KINASE B0464.5 IN CHROMOSOME III (Caenorhabditis elegans)
X52151	Homo sapiens arylsulphatase A mRNA, complete cds.
U22055	Human 100 kDa coactivator mRNA, complete cds.
T47377	S-100P PROTEIN (HUMAN).
T62947	60S RIBOSOMAL PROTEIN L24 (Arabidopsis thaliana)
H09719	TUBULIN ALPHA-6 CHAIN (Mus musculus)
M92287	Homo sapiens cyclin D3 (CCND3) mRNA, complete cds.
U26312	Human heterochromatin protein HP1Hs-gamma mRNA, partial cds.
J02854	MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN);contains element TAR1 repetitive element.
D00860	RIBOSE-PHOSPHATE PYROPHOSPHOKINASE I (HUMAN);.
R54097	TRANSLATIONAL INITIATION FACTOR 2 BETA SUBUNIT (HUMAN).
X86693	H.sapiens mRNA for hevin like protein.
H11084	VASCULAR ENDOTHELIAL GROWTH FACTOR (Cavia porcellus)
X63629	H.sapiens mRNA for p cadherin.
L37792	Human syntaxin 1A mRNA, complete cds.
T86444	PROBABLE NUCLEAR ANTIGEN (Pseudorabies virus)
M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds.
T86749	Human (clone PSK-J3) cyclin-dependent protein kinase mRNA, complete cds.,.
M26383	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds.

Gene ID	Annotations
X54942	H.sapiens ckshs2 mRNA for Cks1 protein homologue.
H16096	MITOCHONDRIAL PROCESSING PROTEASE BETA SUBUNIT PRECURSOR (Rattus norvegicus)
J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds.
H08393	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
X93510	H.sapiens mRNA for 37 kDa LIM domain protein.
U14631	Human 11 beta-hydroxysteroid dehydrogenase type II mRNA, complete cds.
H06524	GELSOLIN PRECURSOR, PLASMA (HUMAN);.
L07648	Human MXI1 mRNA, complete cds.
R44301	MINERALOCORTICOID RECEPTOR (Homo sapiens)
U19969	Human two-handed zinc finger protein ZEB mRNA, partial cds.
J04102	Human erythroblastosis virus oncogene homolog 2 (ets-2) mRNA, complete cds.
L11706	Human hormone-sensitive lipase (LIPE) gene, complete cds.